4I1-3

データ研磨の２部グラフへの適用とTwitterからの意見抽出

Extracted opinions from Twitter using bipartite graph polishing

中原 孝信 *1 Takanobu Nakahara　大内 章子 *2 Akiko Ouchi　宇野 毅明 *3 Takeaki Uno　羽室 行信 *2 Yukinobu Hamuro

*1専修大学 商学部 School of Commerce, Senshu University

*2関西学院大学 経営戦略研究科 Institute of Business and Accounting, Kwansei Gakuin University

*3国立情報学研究所 情報学プリンシプル研究系 Principles of Informatics Research Division, National Institute of Informatics

In this paper, we propose the application of graph polishing method to bipartite graphs to enhance graph features for summarizing the text contents on Twitter related to the keyword ”childcare leave”. First, we extract case frames from tweets to create bipartite graphs of tweets. The proposed ”bipartite graph polishing ” technique is applied to original bipartite graphs to reduce noise and clarify a latent structure. Finally, we enumerate maximal bipartite clique from the polished bipartite graphs to express key topics, and all tweets are then clustered with these topics. Based on the proposed method, we successfully discovered key interesting topics and tweets from 200,000 tweets collected from the year 2012 to 2014.

1. はじめに

育児休業法 (1992 年) が施行されて以来，多くの企業で育児休業制度の導入が進んでいる．2014 年には，育児休業制度が規定されている事業所は，従業員数が 30 人以上の事業所では 94.7%に及び，女性の育児休業取得率は 86.6%になっている [1]．しかしながら，第 1 子の出産を機に有職女性の 54.1%が退職しており [2]，出産・育児を経た就業継続はいまだに困難である．このような中，安倍政権は待機児童数ゼロの実現を掲げたり，女性活躍推進法を新たに制定したり，成長戦略の中核に女性の活用を据えている．日本の労働力を維持するためにも，女性が働きやすい環境を提供し，女性の就業継続率を上昇させることは，重要な課題の１つである．

本研究では，育児休業 (以下，育休) についての Twitter 投稿に注目し，一般の人々の声を要約する方法を提案する．そして，育休に対する率直な意見や，育児と仕事の両立のために必要な政策などの情報を得ることを試みる．2016 年 2 月 15 日に投稿された匿名ブログでは，保育園の入園選考に落ちたことに対して国に不満をぶつけた内容が，子育てをしている母親らの共感を集め，待機児童問題に関して国の政策を動かす程の大きな反響を得ている [3]．SNS やブログでは日々膨大な投稿が行われており，その中に埋もれている重要な意見や，多数の意見を要約して提示することには意義がある．

これまでにも著者らは，安倍首相の育休 3 年の要請という発言 (2013 年 4 月 18 日) によって，ツイートの話題がどのように変化したかを捉える方法を提案した [4]．そこでは，単語間の関係性を表す類似度グラフを構築し，そこから密部分グラフを単語クラスタとして抽出することで，文章要約を実施した．

本研究では，単語間の関係性を一般グラフではなく，格フレームを用いた 2 部グラフで表現し，2 部グラフの研磨手法を適用する．そして，研磨後の 2 部グラフから要素の重複が少ない極大 2 部クリークを抽出し，それらをトピックとして利用する．最終的にそのトピックを含むツイートをクラスタリングすることで文書の要約を実施する．

連絡先: 中原孝信，専修大学，川崎市多摩区東三田 2-1-1，TEL:044-911-0559，nakapara@isc.senshu-u.ac.jp

2. 手法

本研究では，図 1 に示す方法でツイートの文章要約を実施する．まず，(1) ツイートを構文解析し，格助詞句と用言句のペアからなる格フレームを抽出する．そして，格フレームを 2 部グラフで表現する．次に (2)2 部グラフにデータ研磨手法を適用する．データ研磨はグラフのクリーニング方法の 1 つであり，グラフから極大クリークを列挙する際に，同じようなクリークが多数列挙されるという重複問題を解決するためにグラフのクリーニングを実施する．(3) データ研磨後の 2 部グラフから極大 2 部クリークを列挙し，得られた 2 部クリークをトピックとして利用する．そして，(4) そのトピックを含むツイートをクラスタリングすることで要約を行う．最後に比較手法から得られた要約と比較するために，(5) アンケート調査を実施し，提案手法の性能を評価する．

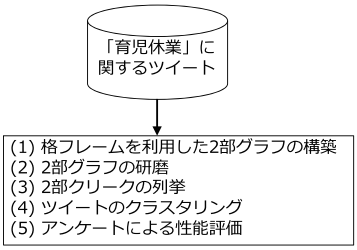

図 1: 分析の概略図

2.1 格フレームを用いた 2 部グラフの構築

これまで文章を表現するために最も利用されてきた方法の 1 つは bag-of-words(BOW) であり，単語の出現だけを考慮したベクトルで文章を表現する方法である．BOW による表現は非常にシンプルで，ときに有用な結果をもたらすが，単語の出現順序や文章の構造を無視しているため，文章の意味を表現する場合にはその点が問題となる．一方で，格フレームは，ガ格やヲ格などの格助詞句と，動詞や形容詞などの用言句のペアによる表現で,「育休を，取得する」「保育園が，一杯だ」など，格フレームによって文章の意味を表すことができる．

本研究では，ツイートから格フレームを抽出するために，日本語の自然言語処理ソフトである KNP[6] を利用し，格解析を実施することで格フレームを抽出する．そして，得られた格フレームを 2 部グラフで表現する．2 部グラフとは，グラフ $G = (V \cup U, E)$ の任意の頂点集合 V と U が枝で接続されたグラフである．抽出した格フレームを構成する格助詞句と用言句をそれぞれ V，U として頂点を枝で結ぶことで 2 部グラフを生成する．

2.2 2 部グラフの研磨

これまで著者らは一般グラフを対象にしたデータ研磨を提案してきた [7]．データ研磨のアイデアは，密度の濃い部分グラフはより濃く，薄い部分グラフはより薄くすることで，本質的な構造を失うことなくグラフを明確化し，列挙されるクリーク数を削減するものである．本研究では，データ研磨を 2 部グラフに対して適用することで，2 部グラフを明確化し，重複の少ない極大 2 部クリークを列挙する．

2 部グラフの頂点集合を $V = \{v_1, v_2, \ldots, v_m\}, U = \{u_1, u_2, \ldots, u_n\}$ とし，$X(v)$ を v に隣接する U の頂点集合とする．また，$X(u)$ を u に隣接する V の頂点集合とする．2 部グラフの研磨アルゴリズムを Algorithm 1 に示す．ここで示すアルゴリズムは，効率の悪い方法ではあるが，理解のし易さを優先させている．

Algorithm 1 2 部グラフ研磨アルゴリズム

```
1: function BIPOLISHING(G = (V ∪ U, E), σ1, σ2)
2:     V, U : 頂点集合, E : 辺集合, σ1, σ2 : 類似度下限値
3:     V', U', E' = φ                          ▷ 頂点集合，辺集合の初期化
4:     for all v ∈ V do
5:         S = φ
6:         for all v' ∈ V do
7:             if sim(X(v), X(v')) ≥ σ1 then▷ 接続関係の類似する頂点を保存
8:                 S = S ∪ {v'}
9:             end if
10:        end for
11:        for all u ∈ U do
12:            if sim(X(u), S) ≥ σ2 then▷ 接続関係が似ていれば枝を張り，似ていなければ張らない
13:                E' = E' ∪ {(v, u)}
14:                V' = V' ∪ {v}
15:                U' = U' ∪ {u}
16:            end if
17:        end for
18:    end for
19:    return G = (V' ∪ U', E')
20: end function
```

2 部グラフの研磨は，接続関係の類似性に着目しており，U への接続関係が指定した閾値 σ_1 以上類似する V の頂点集合を見つけ出し，その頂点集合の各頂点から接続された枝の数が指定した閾値 σ_2 以上になる U を発見する．これは V と U に張られた枝を他の頂点の類似性に基づき新たな 2 部グラフとして再構成している．新たに構成されたグラフを入力として，同様の研磨手法を繰り返し適用し，グラフの構成に変化がなくなるか，もしくはユーザの指定した最大繰り返し回数に達すれば終了する．最終的に得られた 2 部グラフが研磨後の 2 部グラフである．

類似度 (7,12 行目) はさまざまな定義を用いることができるが，本計算では jaccard 係数を利用する．2 つの頂点集合 Y と Z の jaccard 係数による類似度は，式 (1) の通り定義される．

$$sim(Y, Z) = \frac{|Y \cap Z|}{|Y \cup Z|} \tag{1}$$

2.3 極大 2 部クリークの列挙とツイートの要約

格フレームを表した 2 部グラフから密な部分グラフを抽出することで，似た意味を表すクラスタが抽出できていると考え，それをツイート内容の要約に利用する．つまり，研磨後の 2 部グラフから極大 2 部クリークを列挙し，それをトピックとしてツイートの要約を行う．2 部クリークは，G の任意の頂点 $v \in V$ と $u \in U$ に対して枝があるものを 2 部クリークと呼ぶ．そして，ある 2 部クリークが他の 2 部クリークに含まれないとき，その 2 部クリークを極大 2 部クリークと呼ぶ．

データ研磨の特徴の 1 つは，グラフに含まれるノイズを除去し，グラフ構造が明確化されることで，列挙されるクリーク数を大幅に削減できることである．本研究で利用したデータに対しても研磨しなかった場合には，264,733 の極大 2 部クリークが列挙されるが，研磨後の 2 部グラフからの極大 2 部クリーク数は 2,664 で，約 99%の削減ができている．

文章の要約は，ツイートが持つトピック（極大 2 部クリーク）を用いて内容の類似するツイートをクラスタリングして，意味内容が類似したツイートをまとめることで行う．ツイート a に出現するトピック集合を D_a，ツイート b に出現するトピック集合を D_b とすると，2 つのツイートの類似度は以下の式 (2) で計算する．これは式 (1) と同様に jaccard 係数である．

$$jc(a, b) = \frac{|D_a \cap D_b)|}{|D_a \cup D_b|} \tag{2}$$

そして，$jc(a, b)$ がある閾値 σ_3 以上の場合にツイート間に枝を貼り，類似度グラフを生成する．そしてそのグラフに対して Newman クラスタリング [8] を行うことでツイートのクラスタを生成する．

2.4 手法の評価

提案手法では，ツイートをクラスタリングするための素性を 2 部グラフのデータ研磨と極大 2 部クリークによって生成することを示した．ここでは，提案手法の有効性を評価するために異なる 3 つの方法で素性の生成を行う．1 つ目は 2 部グラフの研磨を行わずに極大 2 部クリークを列挙し，それを素性とする方法である．2 つ目は BOW を素性として利用した方法，3 つ目はトピックモデルである Latent Dirichlet Allocation (LDA)[*1] を利用した方法である．

評価の方法は，代表ツイートを一様ランダムに 1 つ選択し，手法毎にそのツイートを含む同一のクラスタから別のツイートをランダムに選択する．そしてアンケート調査を実施し，各手法から選ばれたツイートと代表ツイートを比較してもらい，代表ツイートに最も近いツイートを選んだ手法を 1 位として，4 位までの順位をつけてもらう．複数の手法で甲乙つけがたい場合は同一順位を与えてもらうことにした．

3. 手法の適用

本研究では，2012 年 10 月から 2015 年 1 月 1 日の期間で，「育休」「育児休暇」のどちらかを含むツイートデータ約 28 万件 (13 万ユーザー) を利用した．データクリーニングは，体育休，保育休，教育休を含むツイートの削除と，ストップワードなどを除去し約 20 万ツイートを対象とした．

3.1 2 部グラフ研磨の結果

まず最初に格フレームを抽出し，2 ツイート以上に出現する格フレーム 37,003 種類を分析対象として 2 部グラフを生成した．V は 13,891 種類の格助詞句で U は 4,628 種類の用言句で

*1 計算は R の LDA パッケージを利用した．

あった．この 2 部グラフに対して 2 部グラフの研磨を適用した結果を表 1 に示す．表は σ_1 と σ_2 の値をそれぞれ 0.1 づつ変化させた場合に得られた極大 2 部クリーク数を示している．全体的な傾向は，各閾値を大きくすると得られる極大 2 部クリーク数は少なくなる．これは，σ_1 が大きくなると，用言句への接続関係が強く類似した格助詞句だけが選択されるため，選択される格助詞句が少なくなる．そして σ_2 が大きくなると，選択された格助詞句の多くが共通する用言句への関係を持っていなければ枝が削除されるため疎な 2 部グラフになる．したがって，列挙される極大 2 部クリーク数も少なくなる．

表 1: 研磨の閾値と極大 2 部クリーク数の関係

σ_1	σ_2 0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
0.1	33,999	12,961	8,262	6,635	5,722	4,154	2,896	2,012	1,481
0.2	45,711	14,446	8,658	6,174	4,722	3,168	2,203	1,821	1,482
0.3	29,811	15,126	9,768	6,329	5,372	3,284	2,605	2,367	2,079
0.4	24,922	13,957	9,237	6,368	5,495	3,301	2,803	2,641	2,444
0.5	23,904	13,727	9,201	6,503	5,644	3,400	2,913	2,751	2,554
0.6	20,079	11,768	8,254	5,893	5,425	3,193	2,876	2,781	2,625
0.7	19,768	11,811	8,267	5,915	5,453	3,222	2,911	2,816	2,660
0.8	19,809	11,808	8,272	5,920	5,458	3,226	2,914	2,819	2,663
0.9	19,784	11,810	8,273	5,920	5,458	3,226	2,915	2,820	2,664

3.2 ツイートの要約

本研究では研磨のパラメータとして，最も高い閾値である 0.9 を設定し，得られた 2,664 の極大 2 部クリークを利用してツイートの要約を行った．得られた極大 2 部クリークは，{ 育休 3 年ガ – る, 育休 3 年ガ – 批判噴出, 育休 3 年ガ – 活気, 育休 3 年ガ – 育児子育て支援 } や { 夫ガ – 育児しない, 夫ガ – 育児休業取得する, 夫ガ – 薬局長 } のように比較的意味の取りやすいものが多かった．

これらの極大 2 部クリークを持つツイートをクラスタリングした結果が図 2 である．ツイートの類似度グラフを作成する際に利用した閾値 σ_3 は 0.6 とした．図の点は 1 つのクラスタを示しており，点の大きさは，クラスタに含まれるツイート数に対応している．図には 15 ツイート以上を含むクラスタのみを示している．また，図の軸は各クラスタに含まれるツイートの投稿日から計算した尖度と期間を表しており，縦軸の尖度が高いと短期間で投稿数が多くなっていることを示している．また横軸の期間は，各クラスタで最初に投稿された日から最後に投稿された日までの期間を表しており，同一の話題がどの程度の期間で展開されたかを示している．

例えば，尖度が 66，期間が 280 の「仕事復帰_ヲ-後押しする」というクラスタは全体で 67 件のツイートからなるクラスタで，2013 年 5 月 21 日,22 日に 60 件のツイートが投稿された.「育休 3 年が仕事復帰を後押しするか?」という内容の記事に対しての意見が投稿されたもので,「問題はこの制度の対象にならない非正規雇用が多い点だ」「中小企業が 99.7%の日本では，まず 3 年も待てない」「浦島太郎になりそう」「二人目できたら 6 年休むのか？」「女性にプラスではなく，子供に何がプラスかをまずは考えるべきだ」など否定的な意見が圧倒的に多く,「私は年齢的にぜひ賛成!」という少数の賛成意見もあった．

また，その横の「辞める_デ-なる (否定)」は,「辞めなくなっただけで」という文節に対応したクラスタで,「女性社員，辞めなくなっただけで戦力になっていない．育休，時短の増加で企業疲弊」という記事に対してのツイートが投稿されており，65 件のツイートからなるクラスタで 2013 年 3 月 14,15 日に 57 件の投稿が行われていた.「男性社員にも戦力にならないやつ大量にいるのに」「男女関係なしに，稼げば雇う，稼がないなら解雇でよくないか？」「うちや周りの女性社員は優秀だが」「総合職キャリア組のための施策を腰掛け OL ばかりが使っているから」等の様々な意見が投稿されている．

尖度が低く，期間の長いクラスタとしては,「育休_ガ-取れない」(期間 736, 尖度 0) のクラスタで，育休が取れないことに対しての様々な意見を投稿しており,「補填される額では全然足りないから，簡単に育休って取れないんだよね」「出産するからっていって育休取れない会社なら，いずれにせよ出産の前に辞めるという決断を多くの人がするのでは？」「育休は正直怖くて取れない」「養子だと育休が取れないとか，そんな慣習がこの国に在ったとは.」「こういう職場にいるから普段は感じないけど，仕事で不利益を被る女性は多いんだろうなぁ 気軽に育休取れないとか考えられない」「中小企業や自営業だと育休は取れないよなー」など多岐にわたる意見が投稿されている．

各クラスタはある程度共通のトピックを持ったツイートでまとめられており，トピックがインデックスの役割をすることで，興味のある話題を選択することが可能である．そして，詳細な内容はそのクラスタの各ツイートを確認することで，有益な情報が得られる．

3.3 手法の比較

提案手法との比較のため，2.4 節に示す方法でアンケート調査を実施した．一人の被験者には，10 ツイートを代表ツイートとして選択し，合計で 10 人の被験者にアンケートを実施した．提案手法，研磨なしの極大 2 部クリーク，BOW，LDA の 4 種類の方法から選ばれたツイートと代表ツイートを比較して，内容の近い順に 1 位から 4 位までの順位をつけてもらった．表 2 は，全アンケートの中から 1 つの代表ツイートだけを抜き出したものである．比較 1 は提案手法，2 は BOW，3 は研磨なしの極大 2 部クリーク，4 は LDA である．実際のアンケートでは提示順はランダムで手法もわからないようにした．順位を見ると比較 1 と 3 は同じ内容のツイートで代表ツイートに最も近いと判断されているため 2 つに 1 位が与えられている．

表 2: アンケートの例

方法	ツイート	順位
代表ツイート	そういや、育休だった方が戻ってくるな。2 年ぶりかな？	
比較 1	目安ついたのか〜！育休の方が戻ってくるのかな？	1 位
比較 2	最近ハマってるチョコの名前を検索したらブログ発見＊この製品計画に携わっていた女性社員の方は 1 年半の育休の後またこの製品の担当に戻ってきたみたい！	3 位
比較 3	目安ついたのか〜！育休の方が戻ってくるのかな？	1 位
比較 4	『「パタニティ（父性）・ハラスメント」』／「育休を取得したくてもできなかった」45.5 % 男性の育児参加を阻む「パタハラ」と上司の無理解｜ザ・世論〜日本人の気持ち〜｜ダイヤモンド・オンライン	4 位

表 3 は全代表ツイートに関する結果をまとめたものである．各手法は合計で 100 回評価されており，スコアは，各順位とその頻度の合計を 100 で割った値で，全て 1 位の場合には 1 になる．最も 1 に近い値は提案手法で，続いて研磨なしの極大 2 部クリーク，BOW，LDA の順であった．この結果から提案手法は他の手法よりも意味の類似したツイートでクラスタが構成されているとを示している．提案手法と研磨なしの方法はどち

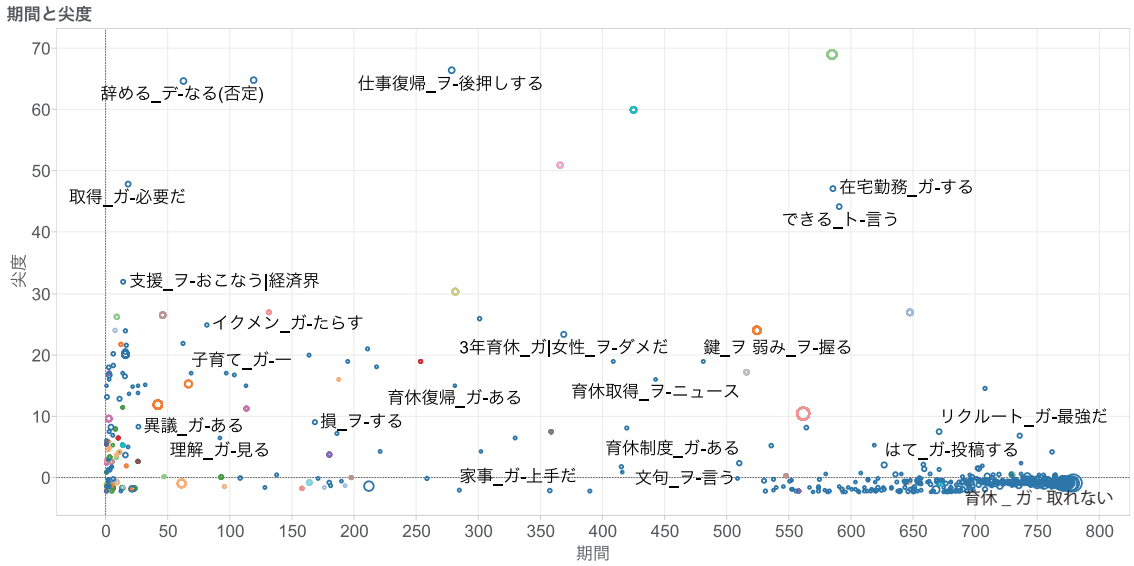

図 2: ツイート要約の結果

らもスコアが 1 に近く僅差であり，研磨なしも比較的意味の類似したツイートでクラスタが構成されている．ただし既に述べたように，素性の数は大きく異なっており，提案手法は 2,664 種類の極大 2 部クリークで，研磨なしは 264,733 種類である．類似したクリークをクラスタリングするためには，素性が多ければ良いわけではなく，2 部グラフの研磨によって，ノイズが除去され重要な格フレームが浮き上がったことによって，有用な結果が得られたと考えられる．

一方で，BOW は格フレームではなく単語 (形態素) の出現のみを扱っており，3,676 種類の形態素を利用している．BOW の場合には形態素のある 1 語が共通することによって，クラスタを構成する場合もあり，素性が細かすぎることで文章の類似性が格フレームに比べて劣っている．LDA は，トピック数を多くしすぎるとトピックの解釈が困難なため，合計で 500 のトピックを生成するようにパラメータを調整したが，素性としては粒度が大きく，意味の類似していないツイートがクラスタリングされてしまった．

これらの結果から，素性の粒度と質が類似するツイートをクラスタリングするためには重要であり，2 部グラフの研磨によって適切な粒度でかつ重要な格フレームが抽出できていると考えられる．

表 3: 手法による比較結果

手法	1 位	2 位	3 位	4 位	スコア
提案手法	86	7	5	2	1.23
研磨なし	79	10	7	4	1.36
BOW	77	10	8	5	1.41
LDA	52	4	7	37	2.29

4. おわりに

本研究では，格フレームを利用した 2 部グラフに，データ研磨を適用することで構造を明確化し，極大 2 部クリークの列挙数が大幅に減少することを示した．また，極大 2 部クリークをトピックとして利用することで，類似するツイートのクラスタリングと有用な情報を抽出できることを示した．育休に関するツイートの要約からは，育休 3 年と仕事復帰に関しては，否定的な意見が多く，育休 3 年という政策は論点がずれていることなど，国民の率直な意見を捉えることができた．今後はクラスタの中で賛成意見，反対意見を評価できるモデルを構築したい．

謝辞

本研究の一部は，科学技術振興機構 CREST，文部科学省の科研費基盤研究 (B)15H03389 の研究助成を受けている．

参考文献

[1] 厚生労働省 (2015)『平成 26 年度雇用均等基本調査』

[2] 第 1 回 21 世紀出生児縦断調査 (平成 22 年出生児) の結果

[3] 日本経済新聞, http://www.nikkei.com/article/DGXZZO76056900T20C14A8000094/

[4] 前川浩基, 内田将史, 大内章子, 宇野毅明, 羽室行信,” データ研磨手法を用いた Twitter ユーザの関係構造変化の検出”, 人工知能学会全国大会論文集, ISSN=”1347-9881”,Vol.28, 2014

[5] M.Berlingerio, D.Koutra, T.Eliassi-Rad, and C.Faloutsos, “ NetSimile: A scalable approach to size-independent network similarity, ” CoRR, vol. abs/1209.2684, 2012.

[6] 黒橋禎夫, 河原大輔,http://nlp.ist.i.kyoto-u.ac.jp/?KNP

[7] 宇野毅明, 中原孝信, 前川浩基, 羽室行信「データ研磨によるクリーク列挙クラスタリング」情報処理学会アルゴリズム研究会報告書, 2014-AL-146(2), pp. 1-8, 2014.

[8] M.E.J.Newman. ”Fast Algorithm for Detecting Community Structure in Networks,” Physical Review E - PHYS REV E, Vol.69, p.066133, 2004.